Principal Component Analysis

How to Simplify and Visualise Data Sets
Plan

- Data sets
- Curse of Dimensionality
- Struggle with Complexity
- Data sets approximation by lines and planes
- Least square definition of mean point
- “Least Square” definition of the first principal component
Plan

- Empirical covariance matrix
- Principal components are eigenvectors of empirical covariance matrix
- PCA scheme
- Eigenfaces and Eigenmuzzles
Principal components analysis (PCA) is a technique used to reduce multidimensional data sets to lower dimensions for analysis. Depending on the field of application, it is also named: (i) the discrete Karhunen-Loève transform, (ii) the Hotelling transform or (iii) proper orthogonal decomposition (POD).
Everybody is a Vector

Here is a dataset

<table>
<thead>
<tr>
<th>age</th>
<th>employment</th>
<th>education</th>
<th>edum</th>
<th>marital</th>
<th>...</th>
<th>job</th>
<th>relation</th>
<th>race</th>
<th>gender</th>
<th>hour</th>
<th>country</th>
<th>wealth</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>State_gov</td>
<td>Bachelors</td>
<td>13</td>
<td>Never_m</td>
<td>...</td>
<td>Adm_cln</td>
<td>Not_in_f</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>51</td>
<td>Self_emp</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Exec_mar</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>13</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>39</td>
<td>Private</td>
<td>HS_grad</td>
<td>9</td>
<td>Divorced</td>
<td>...</td>
<td>Handlers</td>
<td>Not_in_f</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>54</td>
<td>Private</td>
<td>11th</td>
<td>7</td>
<td>Married</td>
<td>...</td>
<td>Handlers</td>
<td>Husband</td>
<td>Black</td>
<td>Male</td>
<td>40</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>28</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Prof_spec</td>
<td>Wife</td>
<td>Black</td>
<td>Female</td>
<td>40</td>
<td>Cuba</td>
<td>poor</td>
</tr>
<tr>
<td>38</td>
<td>Private</td>
<td>Masters</td>
<td>14</td>
<td>Married</td>
<td>...</td>
<td>Exec_mar</td>
<td>Wife</td>
<td>White</td>
<td>Female</td>
<td>40</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>50</td>
<td>Private</td>
<td>9th</td>
<td>5</td>
<td>Married</td>
<td>...</td>
<td>Other_ser</td>
<td>Not_in_f</td>
<td>Black</td>
<td>Female</td>
<td>16</td>
<td>Jamaica</td>
<td>poor</td>
</tr>
<tr>
<td>52</td>
<td>Self_emp</td>
<td>HS_grad</td>
<td>9</td>
<td>Married</td>
<td>...</td>
<td>Exec_mar</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>45</td>
<td>United_S</td>
<td>rich</td>
</tr>
<tr>
<td>31</td>
<td>Private</td>
<td>Masters</td>
<td>14</td>
<td>Never_m</td>
<td>...</td>
<td>Prof_spec</td>
<td>Not_in_f</td>
<td>White</td>
<td>Female</td>
<td>50</td>
<td>United_S</td>
<td>rich</td>
</tr>
<tr>
<td>42</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Exec_mar</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_S</td>
<td>rich</td>
</tr>
<tr>
<td>37</td>
<td>Private</td>
<td>Some_coll</td>
<td>10</td>
<td>Married</td>
<td>...</td>
<td>Exec_mar</td>
<td>Husband</td>
<td>Black</td>
<td>Male</td>
<td>80</td>
<td>United_S</td>
<td>rich</td>
</tr>
<tr>
<td>30</td>
<td>State_gov</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Prof_spec</td>
<td>Husband</td>
<td>Asian</td>
<td>Male</td>
<td>40</td>
<td>India</td>
<td>rich</td>
</tr>
<tr>
<td>24</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Never_m</td>
<td>...</td>
<td>Adm_cln</td>
<td>Own_child</td>
<td>White</td>
<td>Female</td>
<td>30</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>33</td>
<td>Private</td>
<td>Assoc_aci</td>
<td>12</td>
<td>Never_m</td>
<td>...</td>
<td>Sales</td>
<td>Not_in_f</td>
<td>Black</td>
<td>Male</td>
<td>50</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>41</td>
<td>Private</td>
<td>Assoc_voc</td>
<td>11</td>
<td>Married</td>
<td>...</td>
<td>Craft_rep</td>
<td>Husband</td>
<td>Asian</td>
<td>Male</td>
<td>40</td>
<td>*Missing</td>
<td>rich</td>
</tr>
<tr>
<td>34</td>
<td>Private</td>
<td>7th_8th</td>
<td>4</td>
<td>Married</td>
<td>...</td>
<td>Transport</td>
<td>Husband</td>
<td>Amer_Ind</td>
<td>Male</td>
<td>45</td>
<td>Mexico</td>
<td>poor</td>
</tr>
<tr>
<td>26</td>
<td>Self_emp</td>
<td>HS_grad</td>
<td>9</td>
<td>Never_m</td>
<td>...</td>
<td>Farming_f</td>
<td>Own_child</td>
<td>White</td>
<td>Male</td>
<td>35</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>33</td>
<td>Private</td>
<td>HS_grad</td>
<td>9</td>
<td>Never_m</td>
<td>...</td>
<td>Machine</td>
<td>Unmarried</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>38</td>
<td>Private</td>
<td>11th</td>
<td>7</td>
<td>Married</td>
<td>...</td>
<td>Sales</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>50</td>
<td>United_S</td>
<td>poor</td>
</tr>
<tr>
<td>44</td>
<td>Self_emp</td>
<td>Masters</td>
<td>14</td>
<td>Divorced</td>
<td>...</td>
<td>Exec_mar</td>
<td>Unmarried</td>
<td>White</td>
<td>Female</td>
<td>45</td>
<td>United_S</td>
<td>rich</td>
</tr>
<tr>
<td>41</td>
<td>Private</td>
<td>Doctorate</td>
<td>16</td>
<td>Married</td>
<td>...</td>
<td>Prof_spec</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>60</td>
<td>United_S</td>
<td>rich</td>
</tr>
</tbody>
</table>
How transform them into vectors?
Curse of dimensionality

Curse of dimensionality (Bellman 1961) refers to the exponential growth of complexity as a function of dimensionality.

And what to do if dim > 1000?
Two Main Tricks in our Struggle with Complexity

A large space with something interesting inside

Model reduction

Self-simplification in large dim

A “minimal” space with this interesting content

In high dimensionality many different things become similar, if we choose the proper point of view
A 3D representation of an 8D hypercube

The body has the same radial distribution and the same number of vertices as the hypercube.

A very small fraction of the mass lies near a vertex.

Also, most of the interior is void.

(Hamprecht & Agrell, 2002)

(1) In many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the “best-fitting” straight line or plane. Analytically this consists in taking

\[ y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y, \]

or

\[ z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n, \]

where \( y, x, z, x_1, x_2, \ldots x_n \) are variables, and determining the “best” values for the constants \( a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n \) in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books
Data approximation by a straight line. The illustration from Pearson’s paper

\[ \sum_{i} p_i^2 \rightarrow \text{min} \]
The closest approximation =
The widest scattering of projections

1st Principal axis

Maximal dispersion

2nd principal axis
Mean point

\[ \langle \mathbf{X} \rangle = \frac{1}{m} \sum_{i=1}^{m} \mathbf{X}_i \]

\[ \sum_{i=1}^{m} \| \mathbf{X}_i - \langle \mathbf{X} \rangle \|^2 \rightarrow \text{min} \]

\( \mathbf{X}_i \) – datapoints, \( i = 1, \ldots, m \)

\( X_{ij} \) – coordinates of datapoints, \( j = 1, \ldots, n \)
“Least Square” definition of mean point

\[ \Delta^2 = \sum_{i=1}^{m} \| \mathbf{X}_i - \mathbf{Y} \|^2 \to \min, \quad \mathbf{Y} = ? \]

\[ \Delta^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - Y_j)^2 \to \min; \]

\[ \frac{\partial \Delta^2}{\partial Y_j} = -2 \sum_{i=1}^{m} (X_{ij} - Y_j) = -2 \left( \sum_{i=1}^{m} X_{ij} \right) - m Y_j = 0; \]

\[ Y_j = \frac{1}{m} \sum_{i=1}^{m} X_{ij}, \quad \mathbf{Y} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{X}_i = \langle \mathbf{X} \rangle. \]
Centralisation

Let us centralise all data:
Mean Point=The Origin

\[ X_i \mapsto X_i - \langle X \rangle \]
“Least Square” definition of the first principal component

\[ p_i^2 = (X_i - e_1(e_1, X_i), X_i - e_1(e_1, X_i)) \]
“Least Square” definition of the first principal component.

\[ \Delta^2 = \sum_{i=1}^{m} p_i^2 = \sum_{i=1}^{m} (X_i - e_1(e_1, X_i), X_i - e_1(e_1, X_i)) \rightarrow \min; \quad e_1 = ? \]

\[ \Delta^2 = \sum_{i=1}^{m} (X_i - e_1(e_1, X_i), X_i - e_1(e_1, X_i)) = \]

\[ = \sum_{i=1}^{m} (X_i, X_i) - 2\sum_{i=1}^{m} (X_i, e_1)^2 + \sum_{i=1}^{m} (X_i, e_1)^2 = \sum_{i=1}^{m} (X_i, X_i) - \sum_{i=1}^{m} (X_i, e_1)^2; \]

\[ \sum_{i=1}^{m} (X_i, e_1)^2 \rightarrow \max; \quad e_1 = ? \]

**Theorem:** The closest approximation = The widest scattering of projections.
"Least Square" definition of the first principal component.3

**Theorem:** The closest approximation = The widest scattering of projections

\[
\sum_{i=1}^{m} (X_i, e_1)^2 \to \text{max}; \quad e_1 = ?
\]

\[
\sum_{i=1}^{m} (X_i, e_1)^2 = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} X_{ij} e_{1j} \right)^2 = \sum_{i=1}^{m} \left( \sum_{j,k=1}^{n} X_{ij} e_{1j} X_{ik} e_{1k} \right) =
\]

\[
= \sum_{j,k=1}^{n} e_{1j} \left( \sum_{i=1}^{m} X_{ij} X_{ik} \right) e_{1k} = m(e_1, C(X)e_1),
\]

where \( C(X) \) – empirical covariance matrix: \( C(X)_{jk} = \frac{1}{m} \sum_{i=1}^{m} X_{ij} X_{ik} \).
Properties of empirical covariance matrix

\[
C(X)_{jk} = \frac{1}{m} \sum_{i=1}^{m} X_{ij} X_{ik}
\]

1. \(C(X)\) is symmetric: \(C(X)_{jk} = C(X)_{kj}\);

2. \(C(X)\) is positive definite: \((e, C(X)e) \geq 0\).

Indeed, \((e, C(X)e) = \sum_{i=1}^{m} (X_i, e)^2 \geq 0\)

Hence, eigenvalues of \(C(X)\) are non-negative real numbers, \(\lambda_1 \geq \lambda_2 \geq ... \lambda_n \geq 0\)
Principal components are eigenvectors of empirical covariance matrix. 1

\[ C(X)_{jk} = \frac{1}{m} \sum_{i=1}^{m} X_{ij} X_{ik} \]

Eigenvalues of \( C(X) \) are non-negative real numbers, \( \lambda_1 \geq \lambda_2 \geq ... \lambda_n \geq 0 \); \( v_1, v_2, ... v_n \) are the corresponding orthonormal eigenvectors.

We are looking for \( e_1 = \sum_{i=1}^{m} \varepsilon_{1i} v_i \), \( \varepsilon_{1i} = (e_1, v_i) \), \( \sum_{i=1}^{m} \varepsilon_{1i}^2 = 1 \).

\[ C(X)e_1 = \sum_{i=1}^{m} \varepsilon_{1i} C(X)v_i = \sum_{i=1}^{m} \varepsilon_{1i} \lambda_i v_i ; \]

\( (e_1, C(X)e_1) = \sum_{i=1}^{m} \varepsilon_{1i}^2 \lambda_i \rightarrow \text{max under condition} \sum_{i=1}^{m} \varepsilon_{1i}^2 = 1. \)

Let first eigenvalues be different \( \lambda_1 > \lambda_2 > ... \)

In this case, \( \varepsilon_{11}^2 = 1, \varepsilon_{1i} = 0 (i > 1), \ e_1 = \pm v_1 \)
Principal components are eigenvectors of empirical covariance matrix. 2

- Centralise data;
- Subtract projection on the first eigenvector;
- Solve the same minimisation problem again
  - and immediately get: $e_2 = v_2$
- Iterate!
Principal components analysis

- Calculate the empirical mean
- Calculate the deviations from the mean
- Find the covariance matrix
- Find the eigenvectors and eigenvalues of the covariance matrix
- Rearrange the eigenvectors and eigenvalues
- Compute the cumulative energy content for each eigenvector
- Select a subset of the eigenvectors as low-dimensiona basis vectors
- Project the data onto the new basis